Probabilistic Tensor Factorization and Model Selection

Cyril Goutte, Massih-Reza Amini
Interactive Language Technologies
National Research Council Canada
Cyril.Goutte@nrc.ca, Massih-Reza.Amini@nrc.ca

September 30, 2010

Abstract

The well-known formal equivalence between non-negative matrix factorization and multinomial mixture models extends in a fairly straightforward manner to tensors. Among interesting practical implications of this equivalence, this suggests some principled ways to choose the number of factors in the decomposition. We discuss and illustrate two methods to do this in Positive Tensor Factorization.

1 Introduction

Multi-way arrays are a convenient way to represent data with multiple dimensions, such as spatiotemporal, dynamic or multilingual data [12, 8]. Basic techniques for decomposing multi-way arrays are higher-order generalizations of Singular Value Decomposition (SVD), e.g. Tucker or PARAFAC/CANDECOMP [5]. When the data is positive and an additive decomposition is sought, it makes sense to constrain factors to be positive, hence the development of Positive Tensor Factorization (PTF, [10, 12]), the higher-order extension of Non-negative Matrix Factorization (NMF, [9, 7]).

The equivalence between NMF and Probabilistic Latent Semantic Analysis [3], extends in a fairly straightforward manner to provide a probabilistic view of PTF as a mixture of multinomials. In addition to providing new insights into the way factors are estimated, normalized and compared, this equivalence also suggests a natural approach to selecting the appropriate model structure, i.e. the number of factors in the decomposition. Whereas in the context of PTF, this problem is often addressed by overspecifying the model and "hoping" that unnecessary components will vanish, the probabilistic view provides a principled way to deal with this important issue. We will illustrate this on an artificial example with known structure.

2 Probabilistic Tensor Factorization

In order to simplify notation, and without loss of generality, we focus on a third-order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I\times J\times K}$. A factorization of $\boldsymbol{\mathcal{X}}$ is conveniently expressed using the Kruskal operator $[\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!]$, with $\boldsymbol{A} \in \mathbb{R}^{I\times R}$, $\boldsymbol{B} \in \mathbb{R}^{J\times R}$, and $\boldsymbol{C} \in \mathbb{R}^{K\times R}$, such that $\boldsymbol{\mathcal{X}} = [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!]$ means that $x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr}, \forall i, j, k$. For a positive decomposition, $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are positive: this ensures an additive decomposition of $\boldsymbol{\mathcal{X}}$ as a superposition of positive components. In the presence of noise, the factorization is approximate: $\boldsymbol{\mathcal{X}} \approx [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!]$. The quality of the approximation may be measured by different loss functions such as the L2 (quadratic) or a Kullback-Leibler inspired loss [12], which is:

$$KL(\boldsymbol{\mathcal{X}}, [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!]) = \sum_{ijk} x_{ijk} \log \frac{x_{ijk}}{[\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!]_{ijk}} - x_{ijk} + [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!]_{ijk} \tag{1}$$

Assuming that the observations in $\boldsymbol{\mathcal{X}}$ are (or may be transformed to) counts, let us model the probability $P(i, j, k)$ of having an observation in cell (i, j, k) as a mixture of R conditionally independent multinomials: $P(i, j, k) = \sum_r P(r)P(i|r)P(j|r)P(k|r)$. Defining the R-vector $\boldsymbol{\lambda}$ as $\lambda_r = P(r)$, the $I \times R$ matrix $\boldsymbol{A}$ as $a_{ir} = P(i|r)$, the $J \times R$ matrix $\boldsymbol{B}$ as $b_{jr} = P(j|r)$ and the $K \times R$ matrix $\boldsymbol{C}$ as $c_{kr} = P(k|r)$, the third order tensor $\mathcal{P}$ such that $p_{ijk} = P(i,j,k)$, has a positive tensor factorization $\mathcal{P} = [\![\boldsymbol{\lambda}; \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!] \approx \mathcal{X}/N$. Parameters $\theta = (\boldsymbol{\lambda}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ of the mixture of multinomials are typically estimated by maximizing the likelihood over the data $\mathcal{X}$, or equivalently the log-likelihood:

$$L(\theta; \mathcal{X}) = \sum_{ijk} x_{ijk} \log [\![\boldsymbol{\lambda}; \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!]_{ijk} \tag{2}$$

Notice the similarity with Eq. 1, which motivates the probabilistic view of positive tensor factorization:

Proposition 1 (Equivalence) *Given a multi-way array of observations $\mathcal{X}$, the mixture of multinomials $\theta = (\boldsymbol{\lambda}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ that maximizes the likelihood over $\mathcal{X}$ is a minimum-KL Positive Tensor Factorization of $\frac{1}{N}\mathcal{X}$, and any Positive Tensor Factorization of $\mathcal{X}$ also yields a maximum likelihood solution $\widehat{\theta}$ for $\mathcal{P}$.*

Idea of proof: Note that the equivalence of the KL-loss at minimum and the log-likelihood at maximum is not enough as they are optimized over different constraints. However, it is fairly straightforward to show that no PTF is better than the corresponding probabilistic model (the reverse is obvious).

This equivalence has a number of practical implications:

Normalized representation: The probabilistic model is invariant by permutation of the components, as most mixture models, but it removes some of the invariance by rescaling/rotation by forcing factors to be stochastic. In addition, keeping factors on comparable scales makes interpretation easier.

Parameter Estimation: The Expectation-Maximization (EM) algorithm provides a principled parameter estimation technique which guarantees minimization of the objective at each step. Multiplicative updates for non-negative tensor factorization do not guarantee convergence unless appropriate scaling is performed, whereas EM maintains scaling in one step. In addition, Deterministic Annealing helps stabilize the EM solution.

Model Selection: The probabilistic view suggests a number of approaches for selecting the proper number of components in a constructive manner. By contrast, a common way to "select" the number of components in the traditional factorization setting is to overestimate it and trust the parameter estimation procedure with removing useless components by setting the appropriate parameters to zero.

Model selection in the context of PTF corresponds to selecting the right number of components. In mixture models, a quick way to do that is to rely on information criteria [1, 11]. These augment the (log)likelihood by a term penalizing larger models. Although the regularity assumptions for BIC/AIC are known to be incorrect for mixtures of multinomials, they often yield a reasonable idea of the model structure in practice. The simplest and most well-known are [1, 11]:

$$AIC = 2.L(\widehat{\theta}; \mathcal{X}) - 2.P \qquad \text{and} \qquad BIC = L(\widehat{\theta}; \mathcal{X}) - \frac{P}{2} \times \log(N) \tag{3}$$

Two important issues here are the effective number of parameters P and the effective sample size N. In the experiments below, the number of parameters is computed simply as $P = R((I-1)+(J-1)+(K-1)+1)-1$ for the 3D tensor and similarly for other dimensions. In situations such as text where factors are expected to be very sparse, it may make sense to adopt a more realistic estimation of P that does not unduly penalize large models. Similarly, note that the factorization is invariant by scaling of the data tensor $\mathcal{X}$. In order to take that into account, we use the number of cells and not the sum of counts as sample size, ie $N = IJK$.

The second model selection strategy relies on a standard log-likelihood ratio test for nested models. Mixtures of multinomials are naturally nested: if $R_1 < R_2$, the model with R_1 components is a particular case of the model with R_2 components. In that setting, a typical way to test whether two nested models are significantly different is to use a likelihood ratio test. Denoting the two models $[\![\boldsymbol{A}_{R_1}, \boldsymbol{B}_{R_1}, \boldsymbol{C}_{R_1}]\!]$ and $[\![\boldsymbol{A}_{R_2}, \boldsymbol{B}_{R_2}, \boldsymbol{C}_{R_2}]\!]$ (with R_1 and R_2 components, respectively), the (log-)likelihood ratio statistic is computed as:

$$LR = 2 \sum_{ijk} \log x_{ijk} \frac{[\![\boldsymbol{A}_{R_2}, \boldsymbol{B}_{R_2}, \boldsymbol{C}_{R_2}]\!]}{[\![\boldsymbol{A}_{R_1}, \boldsymbol{B}_{R_1}, \boldsymbol{C}_{R_1}]\!]} = 2(L(\theta_{R_2}; \mathcal{X}) - L(\theta_{R_1}; \mathcal{X})) \tag{4}$$

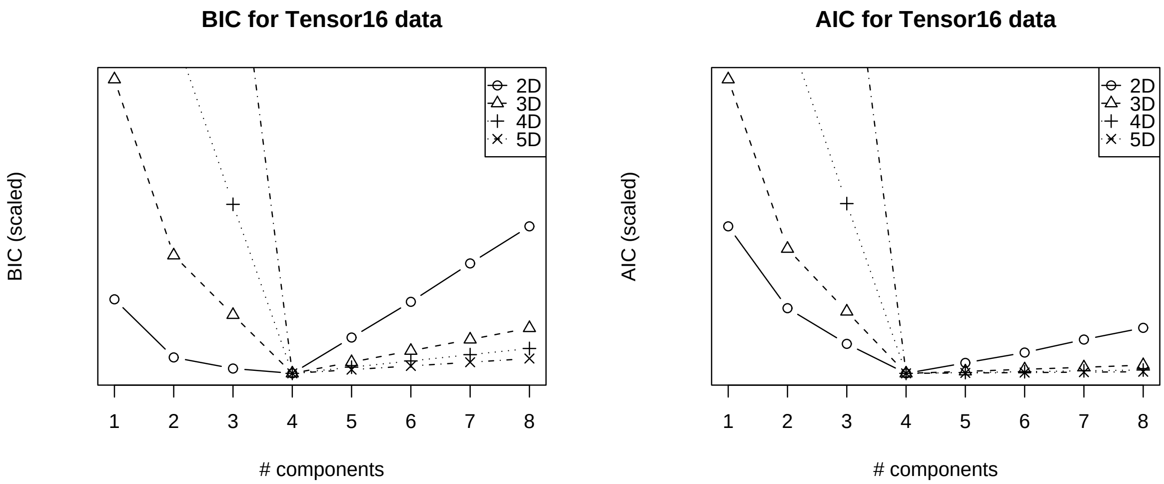

Figure 1: AIC and BIC for Tensors of increasing dimensions (2 to 5), rescaled so that the minimum is at the same level. All suggest correctly 4 components as the right model size.

Under the null hypothesis of no difference between the two models, LR is approximately asymptotically chi-square distributed, with a number of degrees of freedom, df, equal to the difference in the number of parameters in the two models. In the three dimensional case, for example, $df = (R_2 - R_1)(I + J + K - 2)$. This suggests the following model selection strategy: Starting from $R = 1$, a) Estimate model $\widehat{\theta}_R$ with R component; b) Estimate model $\widehat{\theta}_{R+1}$ with $R + 1$ components; c) Repeat previous step as long as $2(L(\widehat{\theta}_{R+1}) - L(\widehat{\theta}_R)) > \chi^2(1 - \alpha, (I + J + K - 2))$. Notice that the number of degrees of freedom between successive models is always the same as it corresponds to adding a single component, so we essentially compare the increase in log-likelihood obtained from each new component to a fixed threshold $\chi^2(1-\alpha, df)$.

3 Experiments

Testing model selection is typically done at either side of a compromise: using artificial data on which the "right" model structure is known, a slightly favorable situation, or using real data, which provides a more realistic setting, but where the correct model structure is unknown. As an illustration, we use the former: we generate several datasets of various dimensions and size, each using 4 factors/components. Factors are overlapping "waves" over either 16 or 24 modalities. We obtain seven datasets: from 16×16 to $16 \times 16 \times 16 \times 16 \times 16$ (1 million cells) and from 24×24 to $24 \times 24 \times 24 \times 24$ tensors.

On each of these 7 arrays, we computed the PTF with $R = 2$ to $R = 8$ components and computed the BIC/AIC criteria and likelihood ratio at each stage. These were estimated using a deterministic annealing variant of EM and multiple random restarts in order to diminish the influence of local minima of the loss function. Figure 1 shows how BIC/AIC behave as the number of components is increased for the 2-to-5-dimensional tensors with 16 modalities. In all cases, the minimum of the information criteria suggest the correct model size of 4 factors. It is apparent that, as expected, BIC penalizes additional components more than AIC. The rate of decrease or increase of the information criteria also varies depending on the data size and dimensionality.

In order to obtain a threshold $\chi^2(1-\alpha, df)$ on the chi-squared statistic, we choose the traditional $\alpha = 0.05$, although other reasonable values of α yield essentially the same result. Depending on the size of the data, and on the difference in degrees of freedom (second row of Table 3), this gives different thresholds, as indicated in the third row of the table. In all cases we see that the observed log-likelihood ratio statistic (lower part of the table) is above the threshold until we reach $R = 4$ factors, and below when we add components beyond that. This again suggests (correctly) that the proper number of factors for these example tensors is 4.

	Tensor16				Tensor24		
Dim:	2	3	4	5	2	3	4
df:	31	46	61	76	47	70	93
$\chi^2(.95, df)$:	44.99	62.83	80.23	97.35	64.00	90.53	116.51
	(log-)likelihood ratio statistic:						
R=2	425.54	6,017.19	26,183.20	81,081.00	945.60	19,242.30	91,841.00
R=3	220.38	2,287.32	18,567.00	111,373.00	459.92	7,025.04	79,887.10
R=4	192.94	2,253.85	18,182.91	128,055.70	440.54	6,721.70	78,848.32
R=5	**15.54**	**22.67**	**36.70**	**49.10**	**23.81**	**36.96**	**42.27**
R=6	**16.09**	**19.45**	**31.39**	**44.20**	**19.47**	**28.65**	**36.68**
R=7	**4.27**	**21.60**	**25.87**	**33.50**	**9.67**	**32.65**	**42.36**
R=8	**10.01**	**15.98**	**35.52**	**42.00**	**16.72**	**28.16**	**54.43**

Table 1: Log-likelihood ratio statistic and χ^2 test. Dim: dataset dimension; df: degrees of freedom; $\chi^2(.95, df)$: test threshold. Values below threshold (bold) indicate that the new component is unnecessary.

Note again that these results were obtained on artificially generated data where the proper model structure is known. The outcome is usually less clear-cut on a real dataset. However these results suggest that these criteria, inspired by the probabilistic model, may at least give a useful indication of a suitable model structure.

4 Conclusion

The probabilistic view of Positive Tensor Factorization is a straightforward extension of the 2D relationship between NMF and PLSA. Nevertheless, it has some interesting implications in terms of normalizing or estimating the factored representation. Using information criteria and likelihood ratio tests, we saw that it is possible to accurately select the number of factors in a situation where the model is complete, ie, there is a model instance that does generate the data. We hope that this relationship can be exploited to further the understanding and help the use of positive tensor factorization.

Acknowledgements: We acknowledge the help of Anicet Choupo with running some of the experiments.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Tr. Automatic Control*, 19(6):716-723, 1974.

[2] W. Buntine. Variational extensions to EM and multinomial PCA. In *ECML'02*, pages 23-34, 2002.

[3] E. Gaussier and C. Goutte. Relation between PLSA and NMF and implications. In *SIGIR*: 601–602, 2005.

[4] C. Goutte, E. Gaussier, K. Yamada. Aligning words using matrix factorisation. In *ACL*: 502–509, 2004.

[5] T. Kolda. Multilinear operators for higher-order decompositions. *T.R. SAND2006-2081*, 2006.

[6] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788-791, 1999.

[7] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *NIPS'13*. MIT Press, 2001.

[8] M. Mørup, L.K. Hansen, J. Parnas, and S.M. Arnfred. Decomposing the time-frequency representation of EEG using non-negative matrix and multi-way factorization. T.R., *Informatics and Mathematical Modeling*, 2006.

[9] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5:111-126, 1994.

[10] P. Paatero. A weighted non-negative least squares algorithm for three-way 'PARAFAC' factor analysis. *Chemometrics and Intel. Lab. Sys.*, 38:223-242, 1997.

[11] G. Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461-464, 1978.

[12] M. Welling and M. Weber. Positive tensor factorization. *Pattern Recognition Letters*, 22(12):1255-1261, 2001.